Lecture14

Saturday, November 6, 2021 4:29 PM

Lecture14

Linear Algebra

Samira Hossein Ghorban
s.hosseinghorban@ipm.ir

Fall, 2021

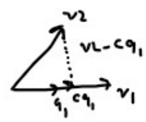
Review: The Gram-Schmidt process

- The Gram-Schmidt process
  1. starts with independent vectors ~~a_1~~, $\dots, v_n$
  2. ends with orthonormal vectors $q_1, \dots, q_n$.
- At step 1: $q_1 = \frac{1}{\|v_1\|} v_1$.
- At step j ($2 \leqslant j \leqslant n$):
  1. it subtracts from a_j its components in the directions $q_1, \dots, q_{j-1}$ that are already settled:
  $$Q_j = v_j - \langle v_j, q_1 \rangle q_1 - \cdots - \langle v_j, q_{j-1} \rangle q_{j-1}.$$
  2. $q_j = \frac{1}{\|Q_j\|} Q_j$.
  3. $\mathrm{span}(\{v_1, \dots, v_j\}) = \mathrm{span}(\{q_1, \dots, q_j\})$.

$Q_2 = v_2 - \langle v_2, q_1 \rangle q_1$

$q_2 = \frac{Q_2}{\|Q_2\|}$

$1 \leq k \leq n \quad \mathrm{span}(\{v_1, \dots, v_k\}) = \mathrm{span}(\{q_1, \dots, q_k\})$



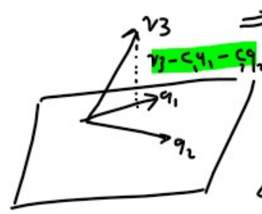
$c = ?$

$\langle v_2 - c q_1, q_1 \rangle = 0$

$\langle v_2, q_1 \rangle - c \underbrace{\langle q_1, q_1 \rangle}_{\|q_1\|^2 = 1} = 0$

$v_2 = c q_1 + v_2 - c q_1$

$\Rightarrow c = \langle v_2, q_1 \rangle$



$c_1, c_2 = ?$

$\langle v_3 - c_1 q_1 - c_2 q_2, q_1 \rangle = 0$

$\langle v_3, q_1 \rangle - c_1 \langle q_1, q_1 \rangle - c_2 \langle q_2, q_1 \rangle = 0$

$\Rightarrow c_1 = \langle v_3, q_1 \rangle$

$\langle v_3 - c_1 q_1 - c_2 q_2, q_2 \rangle = 0$

$\langle v_3, q_2 \rangle - c_1 \langle q_1, q_2 \rangle - c_2 \langle q_2, q_2 \rangle$

$c_2 = \langle v_3, q_2 \rangle$

$\underline{Q_3} = v_3 - \langle v_3, q_1 \rangle q_1 - \langle v_3, q_2 \rangle q_2 \Leftarrow$

$q_3 = \frac{Q_3}{\|Q_3\|}$

The factorization $A = QR$

$A \in M_{m3}$

- The Gram-Schmidt process
  1. starts with independent vectors $a_1, \dots, a_n$, consider these vectors as columns of a matrix A.
  2. ends with orthonormal vectors $q_1, \dots, q_n$, consider these vectors as columns of a matrix Q.
- What is the relation between these matrices A and Q?
- Think about three vectors a_1, a_2, a_3.
- $\mathrm{span}(\{a_1, a_2, a_3\}) = \mathrm{span}(\{q_1, q_2, q_3\})$.
- The idea is to write the a's as combinations of the q's.

$\underline{A} = [a_1 \ \cdots \ a_n] \qquad [q_1, \dots, q_n]$

$u \in \mathrm{span}(\{a_1, \dots, a_k\}) = \mathrm{span}(\{q_1, \dots, q_k\})$

$u = \sum_{i=1}^{n} c_i q_i \qquad \langle q_i, q_j \rangle = 0 \ \text{if} \ i \neq j$

- Think about three vectors a_1, a_2, a_3.
- $\text{span}(\{a_1, a_2, a_3\}) = \text{span}(\{q_1, q_2, q_3\})$.
- The idea is to write the a's as combinations of the q's.

$q_1 \in \text{span}(\{a_1\}) = \text{span}(\{q_1\})$

$$a_1 = \langle a_1, q_1\rangle q_1$$
$$a_2 = \langle a_2, q_1\rangle q_1 + \langle a_2, q_2\rangle q_2$$
$$a_3 = \langle a_3, q_1\rangle q_1 + \langle a_3, q_2\rangle q_2 + \langle a_3, q_3\rangle q_3.$$

$a_3 \in \text{span}(\{a_1, a_2, a_3\}) = \text{span}(\{q_1, q_2, q_3\})$

$u = \sum_{i=1}^{n} c_i q_i \qquad \langle q_i, q_j\rangle = 0 \quad \text{if } i \neq j$

$\langle u, q_j\rangle = \langle \sum_{i=1}^{n} c_i q_i, q_j\rangle = \sum_{i=1}^{n} c_i \langle q_i, q_j\rangle$ ($i \neq j$)

$= c_j \langle q_j, q_j\rangle = c_j$

$c_i = \langle u, q_i\rangle$

$u = \sum_{i=1}^{k} \underbrace{\langle u, q_i\rangle}_{c_i} q_i$

The factorization $A = QR$

- By

$$\begin{aligned} a_1 &= \langle a_1, q_1\rangle q_1 && +0q_2 && +0q_3 \\ a_2 &= \langle a_2, q_1\rangle q_1 && +\langle a_2, q_2\rangle q_2 && +0q_3 \\ a_3 &= \langle a_3, q_1\rangle q_1 && +\langle a_3, q_2\rangle q_2 && +\langle a_3, q_3\rangle q_3; \end{aligned}$$

- we obtain:

$$\underbrace{\begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} q_1 & q_2 & q_3 \end{bmatrix}}_{Q} \underbrace{\begin{bmatrix} \langle a_1, q_1\rangle & \langle a_2, q_1\rangle & \langle a_3, q_1\rangle \\ 0 & \langle a_2, q_2\rangle & \langle a_3, q_2\rangle \\ 0 & 0 & \langle a_3, q_3\rangle \end{bmatrix}}_{R}$$

$A = QR$

$a_1 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \quad a_2 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad a_3 = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}$

QR factorization

- In QR factorization for the matrix A (with independent columns):
  1. the first factor Q has orthonormal columns.
  2. R is upper triangular (The second factor is called R, because the nonzeros are to the right of the diagonal).
- Example:

$\langle a_1, q_1\rangle$

$$A = \begin{bmatrix} 1 & 1 & 2 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} = \begin{array}{c} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \end{bmatrix} \\ \begin{array}{ccc} q_1 & q_2 & \end{array} \end{array} \underbrace{\begin{bmatrix} \sqrt{2} & \frac{1}{\sqrt{2}} & \sqrt{2} \\ 0 & \frac{1}{\sqrt{2}} & \sqrt{2} \\ 0 & 0 & 1 \end{bmatrix}}_{1} = QR$$

$q_1 = \frac{a_1}{\|a_1\|}$

$Q_2 = a_2 - \langle a_2, q_1\rangle q_1$

$q_2 = \frac{Q_2}{\|Q_2\|}$

$Q_3 = a_3 - \langle a_3, q_1\rangle q_1 - \langle a_3, q_2\rangle q_2$

$q_3 = \frac{Q_3}{\|Q_3\|}$

Simple example

- This failure is almost certain when there are several equations and only one unknown:

$$\begin{cases} 2x = b_1 \\ 3x = b_2 \\ 4x = b_3 \end{cases} \qquad x\begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

- In spite of their unsolvability, inconsistent equations arise all the time in practice. They have to be solved!

$A\underline{x} = b \qquad b \in \underline{C(A)}$

$A\hat{x} - b \qquad Ax - b \equiv 0$

$\underline{A\hat{x} - b}$

$3\hat{x} - b_2 \neq 0$

$A\hat{x} \neq b \qquad \underline{A\hat{x} = b}$

$A\hat{x} - b = 0$

- In spite of their unsolvability, inconsistent equations arise all the time in practice. They have to be solved!
- One possibility is to determine x from part of the system, and ignore the rest; this is hard to justify if all m equations come from the same source.
- Rather than expecting no error in some equations and large errors in the others, it is much better to choose the x that minimizes an average error E in all m equations.

$A\hat{x} \neq b$ $\qquad \underline{\underline{Ax}} = \underline{\underline{b}}$

$\underbrace{A\hat{x}} - \underline{b} = 0$

x را $\hat{x}$ به گونه‌ای تعریف بکنیم که $\underline{\|Ax-b\|^2}$ کمینه شود.

Squared error

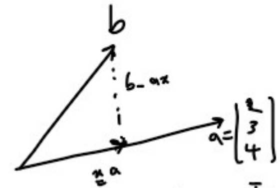
$$\begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix}_A x = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

$$\|Ax-b\|^2 = \left\| \begin{bmatrix} 2x-b_1 \\ 3x-b_2 \\ 4x-b_3 \end{bmatrix} \right\|^2$$

- It is much better to choose the x that minimizes an average error E in the m equations:

$A \to a$

$$\|A\underline{\underline{x}}-b\|^2 = E^2 = \underbrace{(2x-b_1)}^2 + \underbrace{(3x-b_2)}^2 + \underbrace{(4x-b_3)}^2$$



$b = \underline{\underline{ax}} + (b - ax)$

$Pb = ax$

- If there is an exact solution, the minimum error is $E = 0$
- In the more likely case that b is not proportional to a We ***solve*** $Ax = b$ by minimizing

$$E^2 = \|ax-b\|^2.$$

$$\hat{x} = \frac{a^Tb}{a^Ta}$$

Least squares problems with several variables

$2x_1 + 3x_2 = b_1$

$4x_1 - 5x_2 = b_2$

$5x_1 - 6x_2 = b_3$

$$A = \begin{bmatrix} 2 & 3 \\ 4 & -5 \\ 5 & -6 \end{bmatrix}$$

- Now, instead of one column and one unknown x, the matrix now has n columns. The number m of observations is still larger than the number n of unknowns.
- So it must be expected that $Ax = b$ will be inconsistent. Probably, there will not exist a choice of x that perfectly fits the data b.
- In other words, the vector b probably will not be a combination of the columns of A; it will be outside the column space.

$\underline{\underline{b = Ax}}$

- Solution: finding $\hat{x}$ which minimizes $E = \|b - Ax\|$.
- Equivalently, finding $\hat{x}$ such that the error vector $e = b - A\hat{x}$ be perpendicular to that space $C(A)$.

$\hat{x}$ را به گونه‌ای بیابیم که $b - A\hat{x}$ بر $C(A)$ عمود باشد.

Least squares problems with several variables

$b - A\hat{x} \in \underline{\underline{(C(A))}}^\perp \qquad W^\perp$

- Finding $\hat{x}$ such that the error vector $e = b - A\hat{x}$ be perpendicular to that space $C(A)$ is equivalent to

$$A^T(b - A\hat{x}) = 0$$
$$A^TA\hat{x} = A^Tb$$

$N(A^T) = (C(A))^\perp$

$\Rightarrow b - A\hat{x} \in N(A^T)$

$$A^T(b - A\hat{x}) = 0$$

$$A^T A\hat{x} = A^T b$$

- If A^TA is invertible, then the best square estimation :

$$\hat{x} = \underbrace{(A^TA)^{-1}A^T}_{P}\, b.$$

$\Rightarrow b - A\hat{x} \in N(A^T)$

$\Rightarrow A^T(b - A\hat{x}) = 0$

$\Rightarrow A^T b = A^T A\hat{x}$

Matrix A^TA

$A \in M_{mn}(\mathbb{R})$

- The matrix A^TA is certainly symmetric. $A^TA \in M_n$
- $N(A^TA) = N(A)$.
- If A has independent columns, then A^TA is invertible.
- We have shown that the closest point of the column space of A to b is

$$p = A(A^TA)^{-1}A^Tb.$$

- The matrix that gives p is a projection matrix, denoted by P:

$$P = A(A^TA)^{-1}A^T.$$

$x \in N(A) \Rightarrow x \in N(A^TA)$

$Ax = 0 \Rightarrow A^TAx = 0$

$x \in N(A^TA) \Rightarrow A^TAx = 0 \Rightarrow x^TA^TAx = 0$

$\Rightarrow \langle Ax, Ax \rangle = 0 \Rightarrow Ax = 0$

$\Rightarrow x \in N(A)$

$\dim N(A) + \dim C(A) = n$

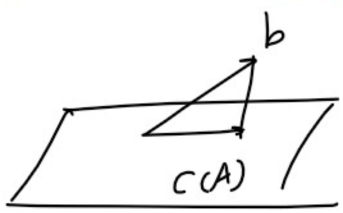
$\dim N(A^TA) + \dim C(A^TA) = n$

$\dim C(A^TA) = n$

Projection matrix

- This matrix has two main properties:

1. P is a symmetric matrix.
2. Its square is itself, again: $P^2 = P$.



$P = A(A^TA)^{-1}A^T$

$P^T = A\left((A^TA)^{-1}\right)^T A = A\left((A^TA)^T\right)^{-1}A^T$

$= A(A^TA)^{-1}A^T = P$

$P^2 = A(A^TA)^{-1}A^TA(A^TA)^{-1}A^T = A(A^TA)^{-1}A^T = P$

Projection matrix

- This matrix has two main properties:
  1. P is a symmetric matrix.
  2. Its square is itself, again: $P^2 = P$.
- Conversely, any symmetric matrix with $P^2 = P$ represents a projection.

$\mathbb{R}^n = C(P) \oplus (C(P))^{\perp}$

$x = Px + \underbrace{(x - Px)}$

$* \ (C(P))^{\perp} = N(P^T) = N(P)$

$* \ P(\underbrace{x - Px}) = Px - \underline{P}^2 x = 0$

Least squares problems with several variables

- So it must be expected that $Ax = b$ will be inconsistent. Probably, there will not exist a choice of x that perfectly fits the data b.
- In other words, the vector b probably will not be a combination of the columns of A; it will be outside the column space.
- Solution: finding $\hat{x}$ which minimizes $E = |\, b - Ax|$.
- Equivalently, finding $\hat{x}$ such that the error vector $e = b - A\hat{x}$ be perpendicular to that space $C(A)$.

Least squares problems with several variables

- Finding $\hat{x}$ such that the error vector $e = b - A\hat{x}$ be perpendicular to that space $C(A)$ is equivalent to

$$A^T(b - A\hat{x}) = 0$$
$$A^T A\hat{x} = A^T b$$

- If $A^T A$ is invertible, then the best square estimation :

$$\hat{x} = (A^T A)^{-1} A^T b.$$

Least-Squares Fitting of Data

- Suppose that we do a series of experiments, and expect the output b to be a linear function of the input t. We look for a straight line $b = C + Dt$.
- To measure the distance to a satellite on its way to Mars!
- We vary the load on a structure, and measure the movement it produces.

$t_0 \longrightarrow b_0$

$t_1 \longrightarrow b_1$

$t_2 \longrightarrow b_2$

$\vdots$

C, D $\quad \boxed{C + Dt = b}$

Example

- Assume three measurements $b_1 = 1, b_2 = 1, b_3 = 3$ for $t = -1, t = 1, t = 2$, respectively.
- Thus, every $C + Dt$ would agree exactly with b:

$$\underbrace{\begin{bmatrix} 1 & -1 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} C \\ D \end{bmatrix}}_{\hat{x}} = \underbrace{\begin{bmatrix} 1 \\ 1 \\ 3 \end{bmatrix}}_{b} \qquad \begin{cases} C - D = 1 \\ C + D = 1 \\ C + 2D = 3 \end{cases}$$

- What is the best estimation?

$A^T A \hat{x} = A^T b$

$\hat{x} = \begin{bmatrix} \hat{C} \\ \hat{D} \end{bmatrix}$

$$\underline{\begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix}} \begin{bmatrix} \hat{C} \\ \hat{D} \end{bmatrix} = \begin{bmatrix} 5 \\ 6 \end{bmatrix}$$

$\hat{C} =$

$\hat{D} =$

$$\begin{cases} C + D t_1 = b_1 \\ \vdots \\ C + D t_m = b_m \end{cases} \Rightarrow \underbrace{\begin{bmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_m \end{bmatrix}}_{A} \begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}$$

$E^2 = \|b - A x\|$ $\qquad \|b - A\hat{x}\| \leq \|b - Ax\|$

$A^T A \hat{x} = A^T b$

$$A^T A = \begin{bmatrix} 1 & \cdots & 1 \\ t_1 & \cdots & t_m \end{bmatrix} \begin{bmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_m \end{bmatrix} = \begin{bmatrix} m & \sum_1^m t_i \\ \sum_1^m t_i & \sum t_i^2 \end{bmatrix}$$

$C e^{-\lambda t} + D e^{-\mu t}$

Weighted Least Squares

- Now suppose that the two observations are not trusted to the same degree.
- A simple least-squares problem is to estimate two observations $x = b_1$ and $x = b_2$.
- if $b_1 \neq b_2$, then we are faced with an inconsistent system of two equations in one unknown:

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} x \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}.$$

$x = b_1$

$x = b_2$

$b_1 \neq b_2$

- Thus

$$E^2 = (x - b_1)^2 + (x - b_2)^2$$

- The best square estimation:

$$\hat{x} = \frac{b_1 + b_2}{2}.$$

$$\begin{cases} C e^{-\lambda t_1} + D e^{-\mu t_1} = b_1 \\ \vdots \\ C e^{-\lambda t_m} + D e^{-\mu t_m} = b_m \end{cases}$$

$$\underbrace{\begin{bmatrix} e^{-\lambda t_1} & e^{-\mu t_1} \\ \vdots & \vdots \\ e^{-\lambda t_m} & e^{-\mu t_m} \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} C \\ D \end{bmatrix}}_{x} = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}$$

$A^T A \hat{x} = A^T b$

$$A^T A = \begin{bmatrix} \sum_{i=1}^{m} (e^{-\lambda t_i})^2 & \sum e^{-\lambda t_i} e^{-\mu t_i} \\ & \sum (e^{-\mu t_i})^2 \end{bmatrix}$$

Weighted Least Squares

$$A^T A = \begin{bmatrix} \sum_{i=1} (e^{\cdots}) & \sum e^{\cdots} \\ \sum & \sum (e^{-\lambda t_i})^2 \end{bmatrix}$$

Weighted Least Squares

- Now suppose that the the value $x = b_1$ may be obtained from a more accurate source from a larger $x = b_2$.
- The simplest compromise is to attach different weights w_1 and w_2 and choose the $\hat{x}_W$ that minimizes the weighted sum of squares:

$$E^2 = w_1(x - b_1)^2 + w_2(x - b_2)^2.$$

- The best square estimation :

$$\hat{x} = \frac{w_1^2 b_1 + w_2^2 b_2}{w_1^2 + w_2^2}.$$

- Generally:

$$(A^T W^T W A)\hat{x}_W = A^T W^T W b.$$

Approximation

A best approximation

A **best approximation** to b by vectors in W is a vector $a in W$ such that

$$\|b - \hat{a}\| \leqslant \|b - a\|$$

for each vector $a \in W$. The vector $\hat{a} \in W$ is called a best approximation to b in W.

Approximation

Theorem

Let W be a subspace of an inner product space V and let b be a vector in V.

(i) *The vector $\hat{a}$ in W is a best approximation to b by vectors in W if and only if $b - \hat{a}$ is orthogonal to every vector in W.*

(ii) *If a best approximation to b by vectors in W exists, it is unique.*

(iii) *If W is finite-dimensional and $\{v_1, \dots, v_n\}$ is any orthonormal basis for W, then the vector*

$$\hat{a} = \sum_{k=1}^{n} \langle b, v_k \rangle v_k$$

is the (unique) best approximation to b by vectors in W.

Thank You!